Multi-Pitch Estimation

Christensen, Mads G.; Stoica, Petre; Jakobsson, Andreas; Jensen, Søren Holdt

Published in:
Signal Processing

DOI (link to publication from Publisher):
10.1016/j.sigpro.2007.10.014

Publication date:
2008

Document Version
Accepted author manuscript, peer reviewed version

Multi-Pitch Estimation

Mads Græsbøll Christensen *,1,2
Dept. of Electronic Systems, Aalborg University, Niels Jernes Vej 12, DK-9220 Aalborg, Denmark

Petre Stoica [3]
Dept. of Information Technology, Uppsala University, P O Box 337, SE-751 05 Uppsala, Sweden

Andreas Jakobsson
Dept. of Electrical Engineering, Karlstad University, Universitetsgatan 2, SE-651 88 Karlstad, Sweden

Søren Holdt Jensen [1]
Dept. of Electronic Systems, Aalborg University, Niels Jernes Vej 12, DK-9220 Aalborg, Denmark

Abstract

In this paper, we formulate the multi-pitch estimation problem and propose a number of methods to estimate the set of fundamental frequencies. The proposed methods, based on the nonlinear least-squares (NLS), MUltiple SIgnal Classification (MUSIC) and the Capon principles, estimate the multiple fundamental frequencies via a number of one-dimensional searches. We also propose an iterative method based on the Expectation Maximization (EM) algorithm. The statistical properties of the methods are evaluated via Monte Carlo simulations for both the single- and multi-pitch cases.

Key words: parameter estimation, estimation theory, pitch estimation, fundamental frequency estimation, periodic signals

1. Introduction

The problem of finding the fundamental frequency, or pitch, of a periodic waveform occurs in many signal processing applications, for example in applications involving speech and audio signals. For instance, in audio processing the fundamental frequency plays a key role in automatic transcription and classification of music [1]. Due to the importance of the problem, a wide variety of fundamental frequency estimation methods have been developed in the literature, e.g., [2–14]. In most cases, these methods are based on a model where only a single set of harmonically related sinusoids are present at the same time. Indeed, the multi-pitch estimation problem, i.e., the problem of estimating the fundamental frequencies of multiple periodic waveforms, is a difficult one, and one that has received much less attention than the single-pitch case, though notable exceptions can be found in [15,1,16,17]. The multi-pitch scenario occurs regularly in music signals, perhaps even more frequently than the single-pitch case, and often also in speech processing. Typically, the situation occurs whenever multiple instruments or speakers are present at the same time or when multiple tones are being played on a musical instrument. The multi-pitch estimation problem can be defined as follows: consider a signal consisting of several, say K, sets of harmonics (hereafter referred to as sources) with fundamental frequencies ω_k, for $k = 1, \ldots, K$, that is corrupted by an additive white complex circularly symmetric Gaussian noise, $w(n)$, having variance σ^2, for $n = 0, \ldots, N-1$, i.e.,

$$x(n) = \sum_{k=1}^{K} \sum_{l=1}^{L} a_{k,l} e^{j\omega_k l n} + w(n), \tag{1}$$

where $a_{k,l} = A_{k,l} e^{j\phi_{k,l}}$, with $A_{k,l} > 0$ and $\phi_{k,l}$ being the amplitude and the phase of the l'th harmonic of the k'th source, respectively. The problem is then to estimate the fundamental frequencies $\{\omega_k\}$, or the pitches, from a set of N measured samples, $x(n)$. In the present work, we assume that the num-

* Contact information: phone: +45 96 35 86 20, fax: +45 96 15 15 83, email: mgc@es.aau.dk

[1] This work was supported by the Intelligent Sound project, Danish Technical Research Council grant no. 26–04–0092

[2] This work was supported by the Parametric Audio Processing project, Danish Research Council for Technology and Production Sciences grant no. 274–06–0521.

[3] This work was supported by the Swedish Science Council (VR).

ber of sources, K, is known and that the number of harmonics, L, of each source is also known and the same for all sources. For the single-pitch case, the order L is often also assumed known in the literature concerned with parametric fundamental frequency estimation (e.g., [6,2,18,15]). Even so, it may seem like a restrictive assumption that K and L are considered to be known, but for many practical applications, it is not required that the order be known precisely. Provided that the order does not vary too much, it is often sufficient to simply assume an average order. The role of an order estimate is mainly to avoid ambiguities in the cost functions that may cause spurious estimates at q/g times the true fundamental frequency (with $q, g \in \mathbb{N}$) such as the well-known problems of halvings and doublings. We note in passing that we here consider the amplitudes and the phases $\{A_{k,l}e^{j\phi_{k,l}}\}$ as nuisance parameters that are not of interest. However, we see from (1) that the complex amplitudes are linear parameters that are, in principle, much easier to find than the nonlinear fundamental frequencies $\{\omega_k\}$. Given the fundamental frequencies $\{\omega_k\}$, the amplitudes and phases can easily be found using one of the estimators proposed in [19]. For more on the topic of sinusoidal amplitude estimation, we refer the interested reader to [19] and the references therein. Furthermore, we remark that also real valued signals can be written using the complex model in (1) through the use of the (downsampled) discrete-time analytic signal [20], provided that there are no harmonics in the real signal near 0 and π relative to N. Here, we have used the complex formulation because of its notational simplicity and because it leads to computationally simpler algorithms.

In this paper, we propose and evaluate a number of estimators for finding the fundamental frequencies $\{\omega_k\}$ based on well-founded principles from statistical signal processing. In particular, we propose an approximate nonlinear least-squares (NLS) method, a MUltiple SIgnal Classification (MUSIC) based method as well as a Capon-based method. These methods have the following simple form:

$$\{\hat{\omega}_k\} = \arg\max_{\{\omega_k\}} \sum_{k=1}^{K} J(\omega_k), \tag{2}$$

where the function $J(\cdot)$ depends only on the source k. This means that an estimate of the set of fundamental frequencies can be obtained by evaluating a cost function $J(\omega_k)$ for a coarse grid of values and then picking the K highest peaks, i.e., costly multi-dimensional searches are avoided. High-resolution estimates can then be found iteratively using the gradients, and in one case also the Hessian, that are derived in this paper for the various cost functions. Additionally, we propose an iterative method based on the Expectation Maximization (EM) principle that is demonstrated to overcome some problems of the NLS method for the multi-pitch case, whereas for the single-pitch case, it is identical to the NLS-based method. We note in passing that if the sources have different numbers of harmonics, the problem becomes somewhat more complicated, but the methods considered here can still be applied. Specifically, the cost function $J(\cdot)$ would have to be calculated for different number of harmonics in order to determine the fundamental frequency, but the fundamental frequencies could still be determined independently for the individual sources.

The rest of the paper is organized as follows: first, in Section 2, we introduce some notation and definitions. In Section 3, we present the proposed multi-pitch estimators along with the assumptions they are based on. Then, in Section 4, we analyze the performance of the estimators using synthetic signals and Monte Carlo simulations. Finally, we conclude the work in Section 5.

2. Preliminaries

We begin by introducing some useful notation, definitions and results. First, constructing a vector formed from M consecutive samples of the observed signal as $\mathbf{x}(n) = [\, x(n) \;\cdots\; x(n+M-1) \,]^T$ with $M \leq N$ and $\mathbf{w}(n) = [\, w(n) \;\cdots\; w(n+M-1) \,]^T$, with $(\cdot)^T$ denoting the transpose, we note that the signal model in (1) can be written as

$$\mathbf{x}(n) = \sum_{k=1}^{K} \mathbf{Z}_k \begin{bmatrix} e^{j\omega_k 1 n} & & 0 \\ & \ddots & \\ 0 & & e^{j\omega_k L n} \end{bmatrix} \mathbf{a}_k + \mathbf{w}(n), \tag{3}$$

where the matrix $\mathbf{Z}_k \in \mathbb{C}^{M \times L}$ has a Vandermonde structure, being constructed from L complex sinusoidal vectors as

$$\mathbf{Z}_k = [\, \mathbf{z}(\omega_k) \;\cdots\; \mathbf{z}(\omega_k L) \,], \tag{4}$$

with $\mathbf{z}(\omega) = [\, 1 \; e^{j\omega} \;\cdots\; e^{j\omega(M-1)} \,]^T$, and $\mathbf{a}_k = [\, a_{k,1} \;\cdots\; a_{k,L} \,]^T$. We note that the constant M is chosen differently in the following sections depending on the method. Next, we define the covariance matrix as

$$\mathbf{R} = \mathrm{E}\left\{\mathbf{x}(n)\mathbf{x}^H(n)\right\}. \tag{5}$$

Here, $\mathrm{E}\{\cdot\}$ and $(\cdot)^H$ denote the statistical expectation and the conjugate transpose, respectively. In practice, the covariance matrix is unknown and is replaced by the sample covariance matrix defined as

$$\widehat{\mathbf{R}} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{x}(n)\mathbf{x}^H(n). \tag{6}$$

Clearly, for $\widehat{\mathbf{R}}$ to be invertible, we require that $M \leq \frac{N}{2}$. In the following, we will assume that M is chosen accordingly whenever the inverse of the covariance matrix is used. For a single source and a high number of samples, i.e., $N \gg 1$, the asymptotic Cramér-Rao lower bound (CRLB) for the k'th source can be shown to be [8]

$$CRLB_k = \frac{6\sigma^2}{N^3 \sum_{l=1}^{L} A_{k,l}^2 l^2}. \tag{7}$$

The CRLB can be seen to depend on the pseudo signal-to-noise ratio (PSNR), defined as

$$PSNR_k = 10 \log_{10} \frac{\sum_{l=1}^{L} A_{k,l}^2 l^2}{\sigma^2} \; [\mathrm{dB}]. \tag{8}$$

Under the assumption that the sources are independent and that the harmonic frequencies are distinct, (7) can also be expected

to hold approximately for the problem of estimating the fundamental frequencies in (1). However, for a low number of samples, the exact CRLB for a fundamental frequency will depend on the parameters of other sources as well.

3. Some Estimators

3.1. *Approximate NLS-based Method*

The first estimator is based on the nonlinear least-squares method. Under the assumption of white Gaussian noise, the NLS method is equivalent to the maximum likelihood method which is well-known to have excellent performance: it attains the CRLB provided that the number of samples is sufficiently high [21,22]. For the sinusoidal estimation problem, the NLS method has been shown to achieve the asymptotic CRLB for large N also in the colored Gaussian noise case [23], and, therefore, the NLS can be expected to be robust to the color of the noise.

For convenience, we define a signal vector containing all N samples of the observed signal as $\mathbf{x} = \mathbf{x}(0)$ with $M = N$. The NLS estimates are obtained as the set of fundamental frequencies and ampltiudes that minimize the 2-norm of the difference between this signal vector and the signal model, i.e.,

$$\{\hat{\omega}_k\} = \arg \min_{\{\mathbf{a}_k\},\{\omega_k\}} \left\| \mathbf{x} - \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{a}_k \right\|_2^2, \tag{9}$$

where $\|\cdot\|_2$ denotes the 2-norm. Assuming that all the frequencies in $\{\mathbf{Z}_k\}$ are distinct and well separated and that $N \gg 1$, (9) can be well-approximated by finding the fundamental frequency of the individual sources, i.e.,

$$\hat{\omega}_k = \arg \min_{\mathbf{a}_k,\omega_k} \|\mathbf{x} - \mathbf{Z}_k \mathbf{a}_k\|_2^2 . \tag{10}$$

Minimizing (10) with respect to the complex amplitudes $\mathbf{a}_k$ gives the estimates $\hat{\mathbf{a}}_k = \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{Z}_k^H \mathbf{x}$, which, when inserted in (10), yields

$$\hat{\omega}_k = \arg \max_{\omega_k} \mathbf{x}^H \mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{Z}_k^H \mathbf{x} \tag{11}$$

$$\approx \arg \max_{\omega_k} \mathbf{x}^H \mathbf{Z}_k \mathbf{Z}_k^H \mathbf{x} \tag{12}$$

where the last line follows from the assumption that $N \gg 1$. Cast in the framework of (2), the resulting cost function is

$$J(\omega_k) = \|\mathbf{Z}_k^H \mathbf{x}\|_2^2, \tag{13}$$

where the matrix product $\mathbf{Z}_k^H \mathbf{x}$ can be implemented efficiently for a linear grid search over ω_k using a fast Fourier transform (FFT) algorithm. The NLS method can be extended to deal with an unknown order for the single-pitch case and colored Gaussian noise in a computationally efficient manner [24]. An alternative interpretation of the approximate NLS estimator is as follows: (13) can be written as $J(\omega_k) = \sum_{l=1}^{L} \|\mathbf{z}(\omega_k l)^H \mathbf{x}\|_2^2$ which is the periodogram power spectral density estimate of $\mathbf{x}$ evaluated at and summed over the harmonic frequencies $\omega_k l$. Furthermore, we note that the NLS cost function in (12) can be written as

$$\|\mathbf{Z}_k^H \mathbf{x}\|_2^2 = \mathrm{Tr}\left[\mathbf{Z}_k^H \mathbf{x}\mathbf{x}^H \mathbf{Z}_k\right]. \tag{14}$$

As an alternative to using the deterministic cost function in (14), we can instead take the expected value after replacing $\mathbf{x}$ by the sub-vector $\mathbf{x}(n)$, with $M < N$, in (14), i.e.,

$$\mathrm{E}\left\{\|\mathbf{Z}_k^H \mathbf{x}(n)\|_2^2\right\} = \mathrm{Tr}\left[\mathbf{Z}_k^H \mathbf{R} \mathbf{Z}_k\right], \tag{15}$$

resulting in the fundamental frequency estimator

$$\hat{\omega}_k = \arg \max_{\omega_k} \mathrm{Tr}\left[\mathbf{Z}_k^H \widehat{\mathbf{R}} \mathbf{Z}_k\right], \tag{16}$$

which instead of matching the signal model to a single snapshot of $\mathbf{x}$ as in (14) matches it to the covariance matrix.

Considering only one source at the time, the gradient of the cost function in (11) can be shown to be

$$\begin{aligned} \nabla J(\omega_k) &\triangleq \frac{\partial J(\omega_k)}{\partial \omega_k} \\ &= \mathbf{x}^H \big[\mathbf{Y}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{Z}_k^H \\ &+ \mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{Y}_k^H \\ &- \mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{W}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k\right)^{-1} \mathbf{Z}_k^H \big] \mathbf{x} \end{aligned} \tag{17}$$

with $\mathbf{Y}_k \in \mathbb{C}^{N \times L}$ being the derivative of the Vandermonde matrix with respect to the fundamental frequency whose elements are defined as

$$[\mathbf{Y}_k]_{nl} \triangleq \left[\frac{\partial}{\partial \omega} \mathbf{Z}_k\right]_{nl} = j(n-1) l e^{j\omega_k l (n-1)}. \tag{18}$$

with $[\mathbf{Y}_k]_{nl}$ denoting the (n,l)'th element of the matrix $\mathbf{Y}_k$. Furthermore, $\mathbf{W}_k \in \mathbb{C}^{L \times L}$ is the derivative of the matrix $\mathbf{Z}_k^H \mathbf{Z}_k$, i.e.,

$$[\mathbf{W}_k]_{lm} \triangleq \left[\frac{\partial}{\partial \omega_k} \mathbf{Z}_k^H \mathbf{Z}_k\right]_{lm} = \sum_{n=0}^{N-1} \left(j(l-m)n\right) e^{j\omega_k (l-m) n}. \tag{19}$$

The gradient in (17) can be used for finding refined estimates. Here, we iteratively find such refined estimates of the fundamental frequency as [4]

$$\hat{\omega}_k^{(i+1)} = \hat{\omega}_k^{(i)} + \delta \nabla J(\hat{\omega}_k^{(i)}), \tag{20}$$

with i being the iteration index and δ a small, positive constant that is found adaptively using approximate line search [25]. For the approximate solution in (12), the corresponding gradient is the much simpler expression

$$\nabla J(\omega_k) \approx 2\,\mathrm{Re}\left(\mathbf{x}^H \mathbf{Y}_k \mathbf{Z}_k^H \mathbf{x}\right). \tag{21}$$

with $\mathrm{Re}(\cdot)$ denoting the real value.

3.2. *MUSIC-based Method*

We proceed to examine a subspace approach based on the MUSIC orthogonality principle (see, e.g., [22,26,27]). In [7,8], it was shown that high resolution fundamental frequency and order estimates can be obtained using this principle, and in

[4] Note that due to the complicated nature of the NLS and Capon-based cost functions, we only use the first order derivative for these, while for the MUSIC cost function, we also derive the Hessian.

[28], the approach was generalized to the multi-pitch estimation problem. We will here briefly review these ideas in the context of this paper, i.e., for the case of known order and number of sources. Assuming that the phases of the harmonics are independent and uniformly distributed on the interval $(-\pi, \pi]$, the covariance matrix and its eigenvalue decomposition (EVD) can be written as

$$\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H = \sum_{k=1}^{K} \mathbf{Z}_k \mathbf{P}_k \mathbf{Z}_k^H + \sigma^2 \mathbf{I}, \tag{22}$$

where $\mathbf{U}$ is formed from the M orthonormal eigenvectors of $\mathbf{R}$, i.e., $\mathbf{U} = [\ \mathbf{u}_1\ \cdots\ \mathbf{u}_M\ ]$, $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues λ_k and

$$\mathbf{P}_k = \mathrm{diag}\left([\ A_{k,1}^2\ \cdots\ A_{k,L}^2\ ]\right). \tag{23}$$

Let $\mathbf{G}$ be the noise subspace formed from the eigenvectors corresponding to the $M - KL$ least significant eigenvalues and note that

$$\mathrm{rank}\left(\sum_{k=1}^{K} \mathbf{Z}_k \mathbf{P}_k \mathbf{Z}_k^H\right) = KL. \tag{24}$$

Then, it can be shown that the noise subspace spanned by $\mathbf{G}$ is orthogonal to all Vandermonde matrices $\{\mathbf{Z}_k\}$ that span the signal subspace formed by the eigenvectors corresponding to the KL most significant eigenvalues. Therefore, the set of fundamental frequencies can be found as [28]

$$\{\hat{\omega}_k\} = \arg\min_{\{\omega_k\}} \sum_{k=1}^{K} \left\|\mathbf{Z}_k^H \mathbf{G}\right\|_F^2, \tag{25}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\mathbf{G}$ is found from the EVD of sample covariance matrix $\widehat{\mathbf{R}}$. Finally, we define the cost function to be maximized for each individual source as [5]

$$J(\omega_k) = -\left\|\mathbf{Z}_k^H \mathbf{G}\right\|_F^2, \tag{26}$$

which can be evaluated efficiently by calculating the FFT of the noise subspace eigenvectors for each segment (see [8] for further details). The gradient of the cost function (26) can be shown to be

$$\nabla J(\omega_k) = -2\,\mathrm{Re}\left(\mathrm{Tr}\left\{\mathbf{Z}_k^H \mathbf{G}\mathbf{G}^H \mathbf{Y}_k\right\}\right), \tag{27}$$

with $\mathbf{Y}_k \in \mathbb{C}^{M\times L}$ having elements defined as in (18). Due to the simplicity of the MUSIC cost function, the Hessian is readily derived as

$$\begin{aligned}\nabla^2 J(\omega_k) &\triangleq \frac{\partial^2 J(\omega_k)}{\partial \omega_k^2} \\ &= -2\,\mathrm{Re}\left(\mathrm{Tr}\left\{\mathbf{Z}_k^H \mathbf{G}\mathbf{G}^H \mathbf{V}_k + \mathbf{Y}_k^H \mathbf{G}\mathbf{G}^H \mathbf{Y}_k\right\}\right)\end{aligned} \tag{28}$$

with $\mathbf{V}_k \in \mathbb{C}^{M\times L}$ being the second order derivative of $\mathbf{Z}_k$, i.e.,

$$[\mathbf{V}_k]_{nl} \triangleq \left[\frac{\partial^2}{\partial \omega_k^2}\mathbf{Z}_k\right]_{nl} = -(n-1)^2 l^2 e^{j\omega_k l(n-1)}. \tag{29}$$

[5] This form is preferred over the more common reciprocal expression due to the ensuing simplicity of the gradient and Hessian.

The gradient and the Hessian can be used for finding refined estimates using Newton's method, i.e.,

$$\hat{\omega}_k^{(i+1)} = \hat{\omega}_k^{(i)} - \delta \frac{\nabla J(\hat{\omega}_k^{(i)})}{\nabla^2 J(\hat{\omega}_k^{(i)})}, \tag{30}$$

The method is initialized for $i = 0$ using the coarse fundamental frequency estimate obtained from (26).

Note that while the NLS method is based on an asymptotic assumption that facilitates finding individual fundamental frequencies independently, there is no such approximation in the MUSIC approach. The covariance matrix decomposition in the MUSIC approach, however, is dependent on the distribution of the phases and the whiteness but not the probabilty density function of the noise. The NLS approach, on the other hand, depends on the noise being Gaussian but it is still asymptotically efficient for colored noise. It should also be noted that, unlike the Capon and NLS approaches, the MUSIC approach requires a priori knowledge about the number of sources for the evaluation of the cost function.

3.3. *Capon-based Method*

We proceed to introduce an estimator based on the Capon approach (see, e.g., [22,29]), which relies on the design of a set of filters that pass power undistorted at specific frequencies, here the harmonic frequencies, while minimizing the power at all other frequencies. Defining the filter bank matrix $\mathbf{H}_k^H$, consisting of L filters of length M, the filter design problem can be stated as the optimization problem:

$$\min_{\mathbf{H}_k} \mathrm{Tr}\left[\mathbf{H}_k^H \widehat{\mathbf{R}} \mathbf{H}_k\right] \quad \text{subject to} \quad \mathbf{H}_k^H \mathbf{Z}_k = \mathbf{I}, \tag{31}$$

where $\mathbf{I}$ is the $L \times L$ identity matrix. The filter bank matrix $\mathbf{H}_k$ solving (31) is given by (see, e.g., [22])

$$\mathbf{H}_k = \widehat{\mathbf{R}}^{-1}\mathbf{Z}_k \left(\mathbf{Z}_k^H \widehat{\mathbf{R}}^{-1} \mathbf{Z}_k\right)^{-1}. \tag{32}$$

This data and frequency dependent filter bank can then be used to estimate the fundamental frequencies by maximizing the power of the filter's output, i.e., $\mathrm{Tr}\left[\mathbf{H}_k^H \widehat{\mathbf{R}} \mathbf{H}_k\right]$. Inserting (32) into this expression yields

$$\hat{\omega}_k = \arg\max_{\omega_k} \mathrm{Tr}\left[\left(\mathbf{Z}_k^H \widehat{\mathbf{R}}^{-1} \mathbf{Z}_k\right)^{-1}\right]. \tag{33}$$

The cost function can be evaluated for different ω_k as

$$J(\omega_k) = \mathrm{Tr}\left[\left(\mathbf{Z}_k^H \widehat{\mathbf{R}}^{-1} \mathbf{Z}_k\right)^{-1}\right]. \tag{34}$$

Similarly to the MUSIC-based method, the computational complexity of the Capon method can be reduced somewhat by calculating $\mathbf{Z}_k^H \widehat{\mathbf{R}}^{-1} \mathbf{Z}_k$ using FFTs. The gradient of the Capon cost function in (34) can be found to be

$$\begin{aligned}\nabla J(\omega_k) = -2\,\mathrm{Re}\Big(\mathrm{Tr}\Big\{&\left(\mathbf{Z}_k^H \widehat{\mathbf{R}}^{-1} \mathbf{Z}_k\right)^{-1} \\ &\times \left(\mathbf{Z}_k^H \widehat{\mathbf{R}}^{-1} \mathbf{Y}_k\right)\left(\mathbf{Z}_k^H \widehat{\mathbf{R}}^{-1} \mathbf{Z}_k\right)^{-1}\Big\}\Big).\end{aligned} \tag{35}$$

The matrix $\mathbf{Y}_k \in \mathbb{C}^{M\times L}$ is constructed as in Section 3.1, i.e., having elements defined as in (18). As in the previous cases, we

iteratively find a refined estimates of the fundamental frequency as

$$\hat{\omega}_k^{(i+1)} = \hat{\omega}_k^{(i)} + \delta \nabla J(\hat{\omega}_k^{(i)}). \tag{36}$$

Alternatively, the filter bank design in (31) can be formulated as the design of a single filter which is subject to L constraints, one for each harmonic. Interestingly, such an approach has some conceptual similarities with the comb-filtering approach of [18].

3.4. *EM-based Method*

Finally, we propose an estimator based on the Expectation Maximization (EM) algorithm [30] (see also [31]). The EM algorithm is an iterative method for maximum likelihood estimation. The method presented here is a special case of [32], which dealt with the estimation of the parameters of superimposed signals. In our case, the superimposed signals are the harmonic sources. We use here the notation of [32]. First, we write the observed signal model in (3) as a sum of K sources in white additive Gaussian noise, i.e.,

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{y}_k \tag{37}$$

where the individual sources are given as

$$\mathbf{y}_k = \mathbf{Z}_k \mathbf{a}_k + \beta_k \mathbf{w}, \tag{38}$$

with the noise source being arbitrarily decomposed into K sources as $\beta_k \mathbf{w}$ where $\beta_k \geq 0$ is chosen so that $\sum_{k=1}^{K} \beta_k = 1$. In the EM algorithm, the set of vectors $\mathbf{y} = \{\mathbf{y}_k\}$ is referred to as the complete data while the observed data $\mathbf{x}$ is referred to as the incomplete data. The complete and incomplete data are related through a many-to-one mapping. The EM algorithm consists of two steps. The first, termed the expectation- or E-step, is the calculation of the conditional expectation of the log-likelihood of the complete data, i.e.,

$$U(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)}) = \int (\ln p_y(\mathbf{y}; \boldsymbol{\theta}))\, p(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}^{(i)}) d\mathbf{y}, \tag{39}$$

where $\boldsymbol{\theta}^{(i)}$ is a vector containing the i'th iteration estimates of the parameters in (3) and $\boldsymbol{\theta}$ is the unknown parameter vector that parameterizes the likelihood function. In the following superscript (i) denotes iteration number. Then, updated parameters are found in the so-called maximization- or M-step by maximizing the above function, i.e.,

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} U(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i)}). \tag{40}$$

For the problem at hand, the two steps of the EM algorithm become particularly simple due to the noise term being Gaussian and white. For details, we refer to [32] and the references therein. Essentially, the E-step reduces to the following where an estimate of the k'th source in noise is obtained based on the parameters of the previous iteration:

$$\hat{\mathbf{y}}_k^{(i)} = \widehat{\mathbf{Z}}_k^{(i)} \hat{\mathbf{a}}_k^{(i)} + \beta_k \left(\mathbf{x} - \sum_{k=1}^{K} \widehat{\mathbf{Z}}_k^{(i)} \hat{\mathbf{a}}_k^{(i)} \right), \tag{41}$$

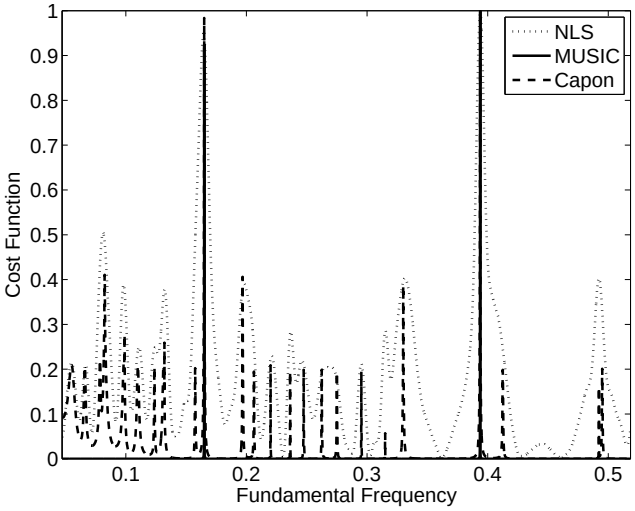

Fig. 1. Example of cost functions (scaled for convenience) for two synthetic sources having five harmonics each and true fundamental frequencies of $\omega_1 = 0.1650$ and $\omega_2 = 0.3937$ for $N = 160$ and $PSNR = 40$ dB.

where $\widehat{\mathbf{Z}}_k^{(i)}$ is the Vandermonde matrix constructed from the fundamental frequency estimate $\hat{\omega}_k^{(i)}$. The problem of estimating the fundamental frequencies then becomes

$$\hat{\omega}_k^{(i+1)} = \arg\max_{\omega_k^{(i+1)}} \hat{\mathbf{y}}_k^{(i)H} \mathbf{Z}_k \left(\mathbf{Z}_k^H \mathbf{Z}_k \right)^{-1} \mathbf{Z}_k^H \hat{\mathbf{y}}_k^{(i)} \tag{42}$$

$$\approx \arg\max_{\omega_k^{(i+1)}} \hat{\mathbf{y}}_k^{(i)H} \mathbf{Z}_k \mathbf{Z}_k^H \hat{\mathbf{y}}_k^{(i)}, \tag{43}$$

and the amplitudes that are needed to form the sources estimates in (41) can be found as

$$\hat{\mathbf{a}}_k^{(i+1)} = \left(\widehat{\mathbf{Z}}_k^{(i+1)H} \widehat{\mathbf{Z}}_k^{(i+1)} \right)^{-1} \widehat{\mathbf{Z}}_k^{(i+1)H} \hat{\mathbf{y}}_k^{(i)}. \tag{44}$$

The M-step in (43) can be seen to be identical to the NLS, with the exception that (43) operates on the estimated source $\hat{\mathbf{y}}_k^{(i)}$ rather than the observed signal $\mathbf{x}$. Accordingly, refined estimates can be obtained in this framework using a gradient reminiscent of the one in Section 3.1. The E-step in (41) and the M-step (43) are then repeated until some convergence criterion is met. As can be seen, the EM algorithm splits the difficult joint estimation problem into a number of much simpler estimation problems by estimating the individual sources. In each iteration of the algorithm, the log-likelihood of the observed data is increased and the algorithm is guaranteed to converge, at least to a local maximum, under mild conditions. The main difficulty in using the EM algorithm is in obtaining the initial parameter estimates required to estimate the individual sources in (41). We here obtain the initial parameters from the approximate NLS.

4. Numerical Results

In this section, we evaluate the performance of the introduced estimators. First we provide an illustrative example based on synthetic signals. Figure 1 shows the cost functions of the proposed estimators, except for the EM-based solution, for a signal of length $N = 160$ consisting of $K = 2$ sources having five unit amplitude harmonics each with $PSNR = 40$ dB. The

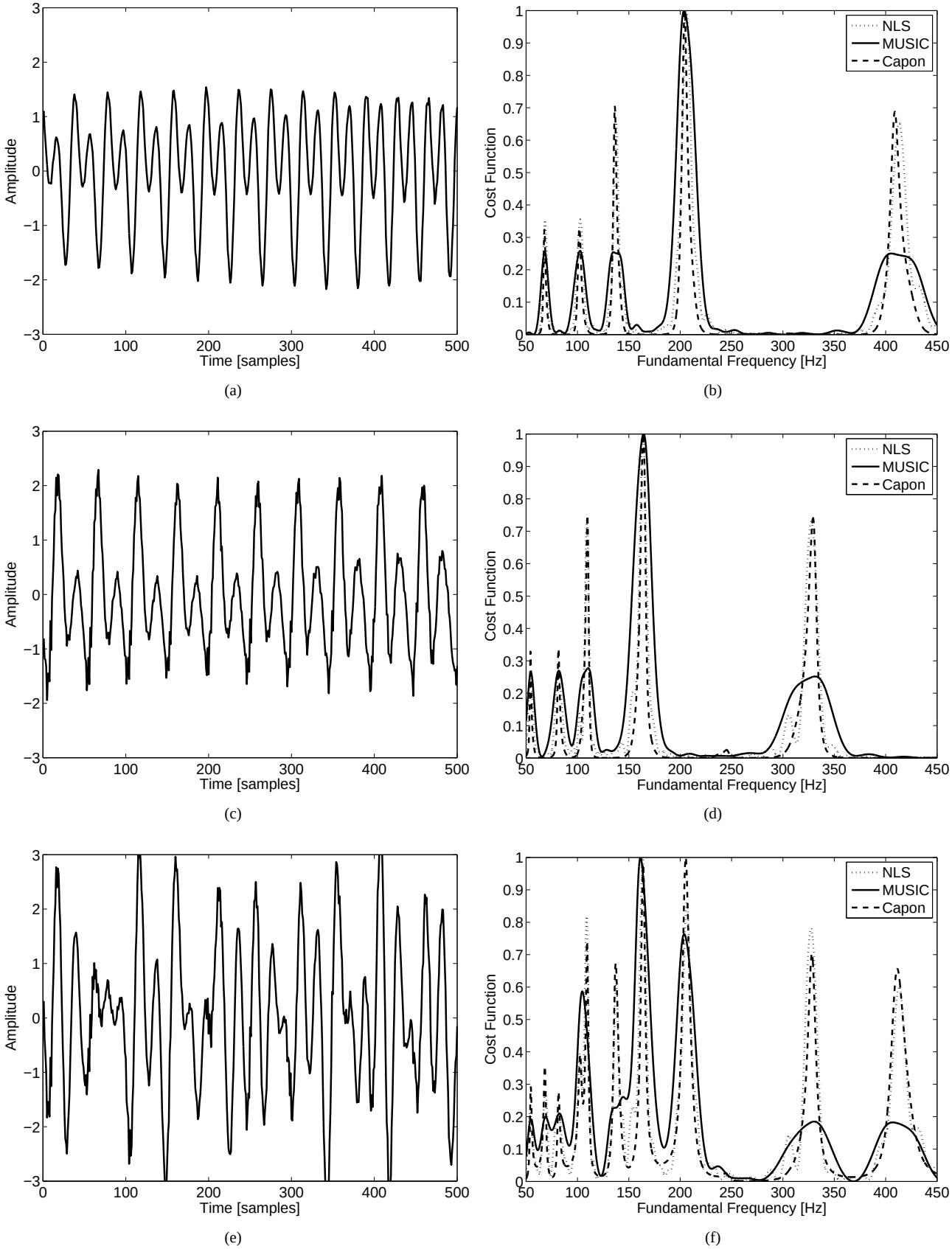

Fig. 2. Example of the proposed methods applied to voiced speech signals. First source (a) and its cost functions (b). Second source (c) and its cost functions (d). Mixture of the two sources (e) and the corresponding cost functions (f).

two sources have fundamental frequencies 0.1650 and 0.3937, respectively. Note that here we show the more traditional reciprocal form of the MUSIC cost function, i.e., $1/\left\|\mathbf{Z}_k^H\mathbf{G}\right\|_F^2$. For the MUSIC-based method, we choose $M = \lfloor N/2 \rfloor$ while for the Capon-based method we used $M = \lfloor 2N/5 \rfloor$. These values were found empirically to result in good performance and will be used in all the experiments reported here [6]. It can be seen that the cost functions have distinct peaks close to the true frequencies with the MUSIC- and Capon-based methods having narrower peaks than the approximate NLS method. Also worth noting is the multi-modal nature of the cost functions with a number of fairly sharp false peaks. Indeed, the multitude of local maxima shows why the fundamental frequency estimation problem is a difficult one. At first sight, this appears to be less of an issue for the MUSIC-based approach, but upon closer inspection, it can be observed that MUSIC generally suffers from this problem too.

The next example is based on two real voiced speech signals sampled at 8 kHz. We have plotted the time signals of the two sources with $N = 500$ in Figures 2(a) and Figures 2(c) and the sum of these sources in Figure 2(e) with the figure showing a more complicated signal. The corresponding cost functions are depicted in Figures 2(b), 2(d) and 2(f). It can be seen, that all the methods correctly identify the fundamental frequencies of the individual sources. But it can also be seen that the cost functions contain a number of spurious peaks. In Figure 2(f), the cost functions are even more complicated, but the methods are still able to find the fundamental frequencies of the two sources.

We proceed to evaluate the proposed estimators using Monte Carlo simulations by generating signals according to the model (1) with the phases and the noise being randomized over realizations. For all combinations of parameters 100 Monte Carlo trials are run. The estimators are evaluated for two sources having fundamental frequencies, $\omega_1 = 0.1580$ and $\omega_2 = 0.6364$, and with $L = 3$, and for one harmonic source of 0.6364. Note that this case is somewhat more difficult that that shown in Figure 1 due to the near-intereger relation between the two fundamental frequencies. We compare the root mean square estimation error (RMSE) of the estimators and the asymptotic CRLB given in (7). Here, the RMSE is defined as

$$RMSE = \sqrt{\frac{1}{SK}\sum_{k=1}^{K}\sum_{s=1}^{S}\left(\hat{\omega}_k^{(s)} - \omega_k\right)^2}, \qquad (45)$$

with ω_k and $\hat{\omega}_k^{(s)}$ being the true fundamental frequency and the estimate, respectively, and with S being the number of Monte Carlo trials. The RMSE is calculated jointly for both sources. We test two different cases for the amplitudes, namely one where all amplitudes are set to unity, i.e., $A_{k,l} = 1,\ \forall k, l$, and one where the amplitudes of each source are decaying, as could be expected for natural spectra, here exemplified by $A_{k,l} = 1/l$. The fundamental frequency estimates are obtained in each Monte Carlo simulation as follows: First, the cost functions (13), (26), and (34) are evaluated on a coarse grid. Then, these coarse estimates are used to initialize the gradient-based methods that are used to obtain refined estimates. For the MUSIC- and Capon-based methods, the gradients of (26) and (34) are used, whereas for NLS method, the gradient for the approximate cost function (13) was found not to produce high resolution estimates. Instead, the gradient was derived for this case based on (11). For the EM algorithm, we used initial parameter estimates from the NLS estimator to form the source estimates with $\beta_k = \frac{1}{K},\ \forall k$. Then, the NLS estimator is applied to each of these sources using the approximate NLS cost function in (13) and to initialize the gradient-based method. A mere 10 iterations of the EM algorithm were found to be sufficient for the application at hand. We note that the NLS cost function in (11) is approximate, being based on neglecting the inner products between the sources. Also, for one harmonic source, the NLS method is exact, meaning that there is no approximation in the estimate (11). Moreover, the NLS and the EM methods are identical for the single-pitch case.

We start out by presenting the results for the unit amplitude case. The RMSEs are shown in Figures 3(a) and 3(b) as a function of N for one and two sources, respectively. Similarly, in Figures 3(c) and 3(d), the RMSEs are shown as a function of the $PSNR$ for one and two sources. It can be seen that for the case of one harmonic source, all estimators perform well, for all tested PSNRs with NLS having the best performance. For two sources, however, the RMSE of the NLS method performs poorly while both the MUSIC- and Capon-based methods follow the CRLB closely. The EM algorithm can be seen to have excellent performance attaining the CRLB. It can also be observed that all methods exhibit thresholding effects below 10 dB while the NLS method appears to saturate at PSNRs above 20 dB.

Next, we consider the case where the sinusoidal amplitudes are decaying. It is not clear from the theoretical derivations how this should affect the performance of the estimators. Therefore this is investigated in simulations similar to those in the previous section. The results are shown in Figures 3(a) and 3(b) as the RMSE as a function of N for one and two sources, respectively. Similarly, in Figures 3(c) and 3(d), the RMSEs are shown as a function of the $PSNR$ for one and two sources. The general conclusions that can be drawn from these figures are the same as those from Figure 3. However, comparing Figures 3 and 4, a number of peculiarities can be noted. First of all, the decaying amplitudes appear to cause a larger gap between the RMSE and the CRLB for the methods that are based on the covariance matrix, namely MUSIC and Capon, while the performance of the NLS and EM methods is unaffected by this. More importantly, the threshold below which the RMSEs differ from the CRLB by an order of magnitude can now be seen from Figure 4(d) to be different for the various methods. It now appears that the MUSIC-based method is more sensitive to noise than the Capon and EM methods.

In a final experiment, the RMSE is studied as a function of the difference between the fundamental frequencies of two harmonic sources, i.e., $\Delta = |\omega_1 - \omega_2|$, for a PSNR of 40 dB and $N = 160$. The results are shown in Figures 5(a) and 5(b)

[6] The reason for having different values of M for the two methods is that they exhibit different sensitivity to the choice of M.

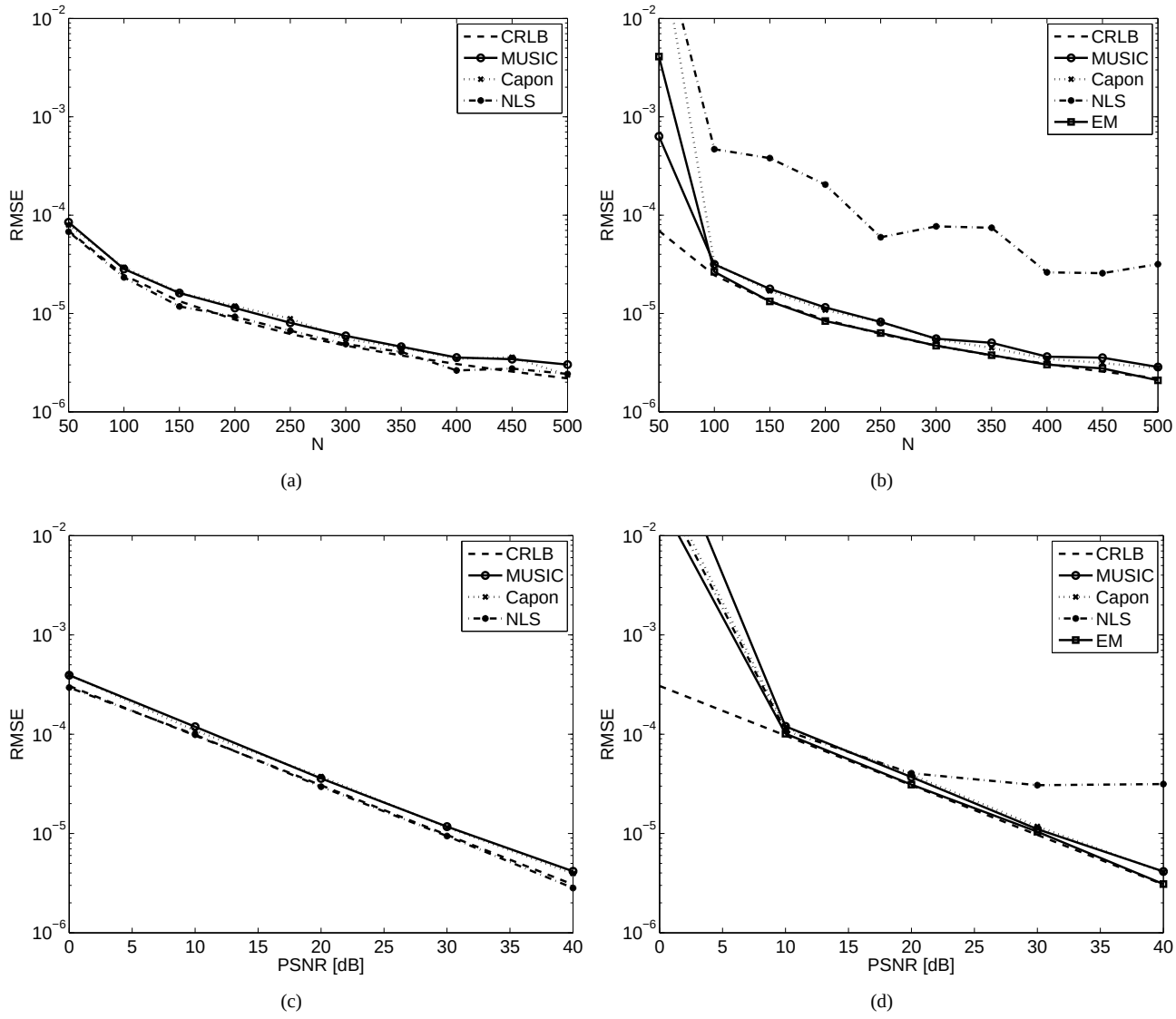

Fig. 3. Estimated RMSEs for unit amplitudes. RMSE as a function of N for $PSNR = 40$ dB for one (a) and two (b) sources. RMSE as a function of $PSNR$ for $N = 400$ for one (c) and two (d) sources.

for unit and decaying amplitudes, respectively. It can be seen that the EM algorithm performs the best for closely spaced harmonics and that the approximate NLS method performs the worst. The Capon-based approach can be observed to be slightly worse than the EM algorithm for unit amplitudes but it still outperforms the MUSIC-based method. For this experiment, we used 100 iterations in the EM algorithm. The decaying amplitudes can be seen to cause a degradation of performance of the methods compared to the unit amplitudes.

At first sight, the conditions of the simulations reported here may seem overly simplisitic. Indeed, one would expect speech and audio signals to contain many sources and many harmonics. However, for more and more sources and harmonics the experiments will be become increasingly complicated and difficult to analyze and make sense of. As the number of sources grow, the interaction effects between the different sources will only become worse, thereby degrading performance of the estimators. As we have seen it is, though, still possible to make some interesting and useful observations from the results presented here. For example, some important properties of the estimators can be determined like efficiency and consistency. Also, our experiments show that the proposed estimators exhibit different sensitivities to differences in frequency, the amplitude distribution and different thresholding effects. These are all very useful observations. Specifically, it appears that the EM and Capon methods are the most robust and are better able to resolve closely spaced harmonic sets.

When making a choice between the various estimators, the complexity should also be taken into consideration. The Capon- and MUSIC-based approaches both have complexity $\mathcal{O}(N^3)$ (since M is proportional to N) due to the matrix inversions, the matrix products and the EVD. On the other hand, the NLS approximation based on the FFT, and thus also the EM algorithm, has complexity $\mathcal{O}(N \log N)$ assuming that the FFT size is chosen proportionally to N. However, the NLS gradient that was used has complexity $\mathcal{O}(N^3)$ due to the matrix inversions,

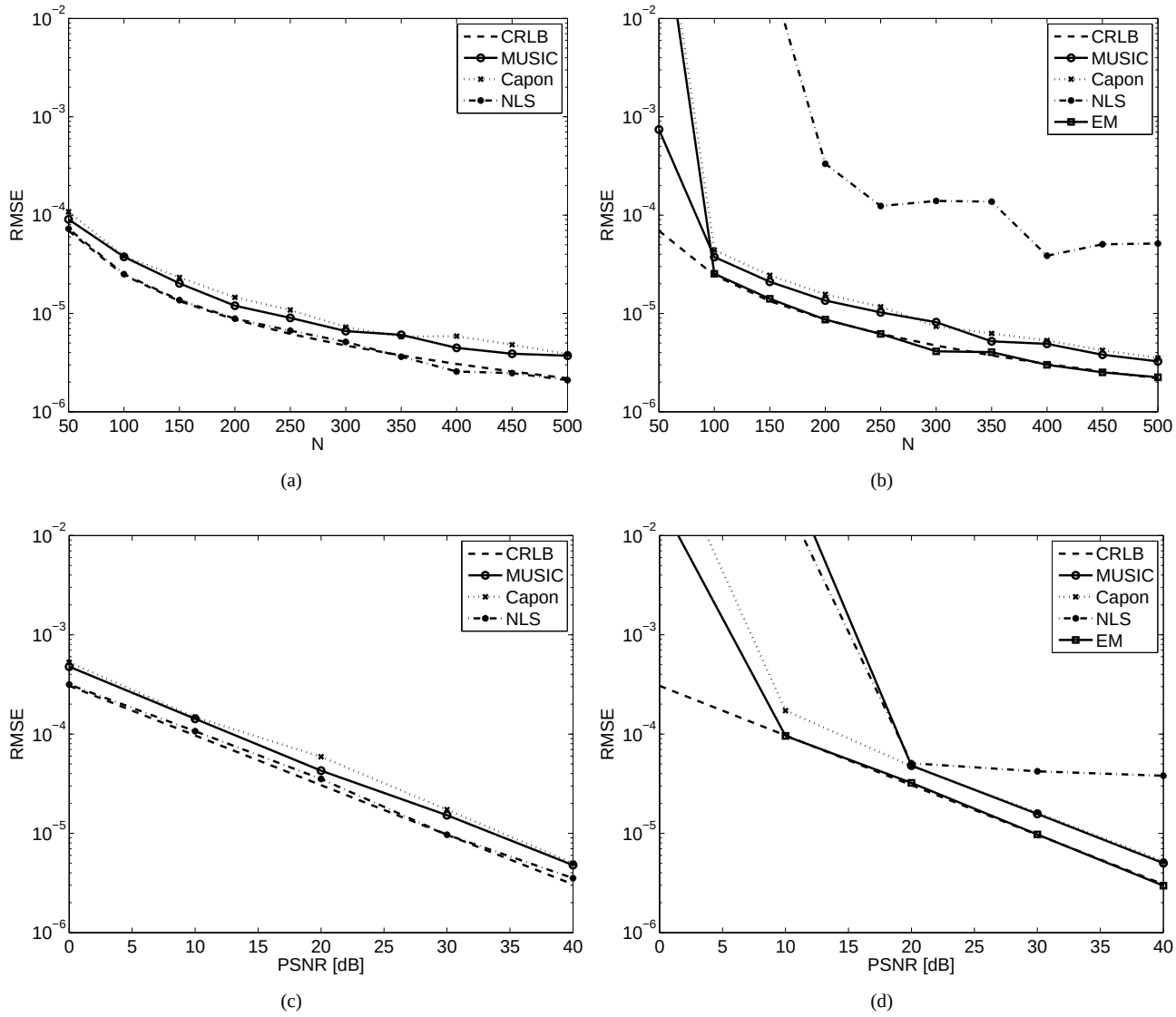

Fig. 4. Estimated RMSEs for decaying amplitudes. RMSE as a function of N for $PSNR = 40$ dB for one (a) and two (b) sources. RMSE as a function of $PSNR$ for $N = 400$ for one (c) and two (d) sources.

as does the expression in (11). If refined estimates are not desired, this seems to favor the EM-based estimator for complexity constrained situations such as real-time processing of speech and audio signals. Also, considering that the noise may very well be colored in some applications and that the NLS, and thus also the EM algorithm, is still asymptotically efficient for colored noise, this is yet another argument that favors the EM algorithm.

5. Conclusions

We have considered the problem of estimating the fundamental frequencies of superpositions of periodic waveforms, also known as the multi-pitch estimation problem. We have proposed a number of estimators that are based on one-dimensional evaluations of cost functions, namely the approximate nonlinear least-squares (NLS), MUSIC- and Capon-based techniques. Additionally, we have also proposed an iterative method based on the expectation maximization (EM) algorithm, which is identical to the NLS method for the single pitch case, and consists of a number of independent NLS estimators for the multi-pitch case. The basic assumptions for these methods to work for the multi-pitch estimation problem have been outlined and their finite sample performance has been evaluated using Monte Carlo simulations. It has been found that the MUSIC- and Capon-based methods have good statistical performance for both the multi- and single-pitch cases, following the Cramér-Rao lower bound (CRLB) closely. As expected, the approximate NLS has excellent performance for the single-pitch case but does not perform well for the multi-pitch case. The EM algorithm is able to mitigate the shortcomings of the NLS for the multi-pitch case as it was found to have excellent performance attaining the CRLB for the number of observations considered here. For closely spaced fundamental frequencies and decaying amplitudes, the Capon approach has been found to have a performance superior to that of the MUSIC method and the EM

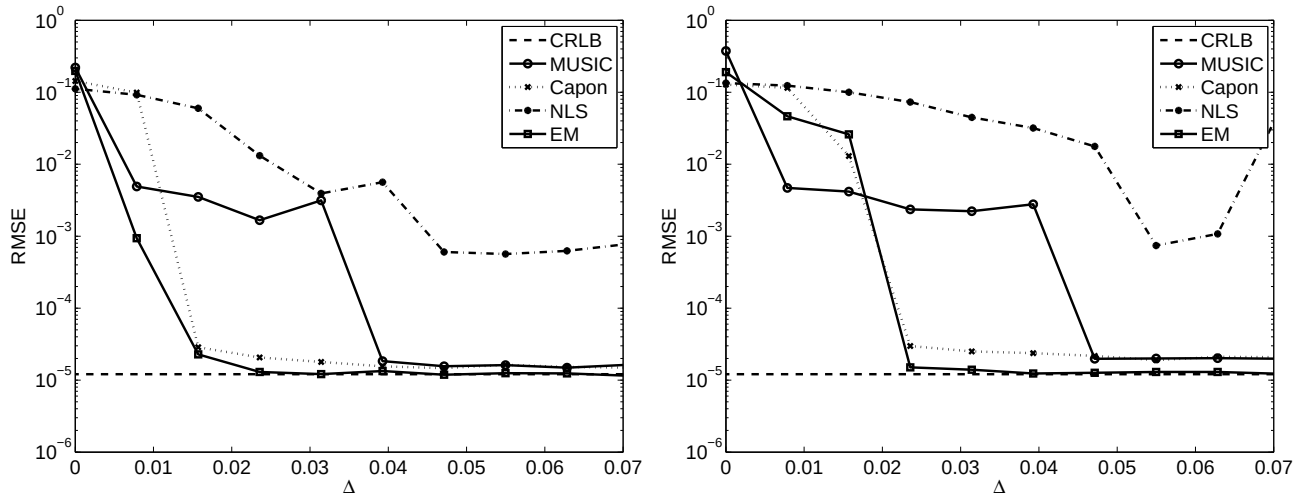

Fig. 5. RMSE versus the difference between the fundamental frequencies of two sources with $N = 160$ and $PSNR = 40$ dB for unit amplitudes (a) and decaying amplitudes (b).

algorithm once again outperformed the other estimators.

References

[1] A. Klapuri, M. Davy (Eds.), Signal Processing Methods for Music Transcription, Springer, New York, 2006.

[2] H. Li, P. Stoica, J. Li, Computationally efficient parameter estimation for harmonic sinusoidal signals, Signal Processing 80 (2000) 1937–1944.

[3] M. Davy, S. Godsill, J. Idier, Bayesian analysis of western tonal music, J. Acoust. Soc. Am. 119(4) (2006) 2498–2517.

[4] S. Godsill, M. Davy, Bayesian computational models for inharmonicity in musical instruments, in: Proc. IEEE Workshop on Appl. of Signal Process. to Aud. and Acoust., 2005, pp. 283–286.

[5] S. Godsill, M. Davy, Bayesian harmonic models for musical pitch estimation and analysis, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Vol. 2, 2002, pp. 1769–1772.

[6] K. W. Chan, H. C. So, Accurate frequency estimation for real harmonic sinusoids, IEEE Signal Processing Lett. 11(7) (2004) 609–612.

[7] M. G. Christensen, S. H. Jensen, S. V. Andersen, A. Jakobsson, Subspace-based fundamental frequency estimation, in: Proc. European Signal Processing Conf., 2004, pp. 637–640.

[8] M. G. Christensen, A. Jakobsson and S. H. Jensen, Joint high-resolution fundamental frequency and order estimation, IEEE Trans. on Audio, Speech and Language Processing 15(5) (2007) 1635–1644.

[9] W. Hess, Pitch Determination of Speech Signals, Springer-Verlag, Berlin, 1983.

[10] W. Hess, Pitch and voicing determination, in: S. Furui, M. M. Sohndi (Eds.), Advances in Speech Signal Processing, Marcel Dekker, New York, 1992, pp. 3–48.

[11] A. de Cheveigné, H. Kawahara, YIN, a fundamental frequency estimator for speech and music, J. Acoust. Soc. Am. 111(4) (2002) 1917–1930.

[12] B. Resch, M. Nilsson, A. Ekman, W. B. Kleijn, Estimation of the instantaneous pitch of speech, IEEE Trans. on Audio, Speech and Language Processing 15(3) (2007) 813–822.

[13] A. Kurshid, S. L. Denham, A temporal-analysis-based pitch estimation system for noisy speech with a comparative study of performance of recent systems, IEEE Trans. Neural Networks 14(4) (2004) 112–1124.

[14] D. Charalampidis, V. B. Kura, Novel wavelet-based pitch estimation and segmentation of non-stationary speech, in: 8th International Conference on Information Fusion, Vol. 2, 2005.

[15] R. Gribonval, E. Bacry, Harmonic Decomposition of Audio Signals with Matching Pursuit, IEEE Trans. Signal Processing 51(1) (2003) 101–111.

[16] A. Klapuri, Multiple fundamental frequency estimation based on harmonicity and spectral smoothness, IEEE Trans. Speech and Audio Processing 11(6) (2003) 804–816.

[17] S. S. Abeysekera, Multiple pitch estimation of poly-phonic audio signals in a frequency-lag domain using the bispectrum, in: Proc. IEEE Int. Symp. Circuits and Systems, Vol. 14(4), 2004, pp. 469–472.

[18] A. Nehorai, B. Porat, Adaptive comb filtering for harmonic signal enhancement, IEEE Trans. Acoust., Speech, Signal Processing 34(5) (1986) 1124–1138.

[19] P. Stoica, H. Li, J. Li, Amplitude estimation of sinusoidal signals: Survey, new results and an application, IEEE Trans. Signal Processing 48(2) (2000) 338–352.

[20] S. L. Marple, Computing the discrete-time "analytic" signal via FFT, IEEE Trans. Signal Processing 47 (1999) 2600–2603.

[21] S. M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice-Hall, 1993.

[22] P. Stoica, R. Moses, Spectral Analysis of Signals, Pearson Prentice Hall, 2005.

[23] P. Stoica, A. Jakobsson, J. Li, Cisiod parameter estimation in the coloured noise case: Asymptotic Cramér-Rao bound, maximum likelihood, and nonlinear least-squares, in: IEEE Trans. Signal Processing, Vol. 45(8), 1997, pp. 2048–2059.

[24] M. G. Christensen, S. H. Jensen, Variable order harmonic sinusoidal parameter estimation for speech and audio signals, in: Rec. Asilomar Conf. Signals, Systems, and Computers, 2006.

[25] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[26] R. O. Schmidt, Multiple emitter location and signal parameter estimation, IEEE Trans. Antennas Propagat. 34(3) (1986) 276–280.

[27] G. Bienvenu, Influence of the spatial coherence of the background noise on high resolution passive methods, in: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 1979, pp. 306–309.

[28] M. G. Christensen, A. Jakobsson, S. H. Jensen, Multi-pitch estimation using harmonic MUSIC, in: Rec. Asilomar Conf. Signals, Systems, and Computers, 2006.

[29] J. Capon, High-resolution frequency-wavenumber spectrum analysis, Proc. IEEE 57(8) (1969) 1408–1418.

[30] N. M. Laird, A. P. Dempster, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Ann. Roy. Stat. Soc. (1977) 1–38.

[31] P. Stoica, Y. Selen, Cyclic minimizers, majorization techniques, and the expectation-maximization algorithm: a refresher, IEEE SP Mag. 21(1) (2004) 112–114.

[32] M. Feder, E. Weinstein, Parameter estimation of superimposed signals using the EM algorithm, IEEE Trans. Acoust., Speech, Signal Processing 36(4) (1988) 477–489.